Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks

Year: 2022 | Citations: 466 | Authors: Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, Aniruddha Kembhavi

---

Abstract

We propose Unified-IO, a model that performs a large variety of AI tasks spanning classical computer vision tasks, including pose estimation, object detection, depth estimation and image generation, vision-and-language tasks such as region captioning and referring expression, to natural language processing tasks such as question answering and paraphrasing. Developing a single unified model for such a large variety of tasks poses unique challenges due to the heterogeneous inputs and outputs pertaining to each task, including RGB images, per-pixel maps, binary masks, bounding boxes, and language. We achieve this unification by homogenizing every supported input and output into a sequence of discrete vocabulary tokens. This common representation across all tasks allows us to train a single transformer-based architecture, jointly on over 90 diverse datasets in the vision and language fields. Unified-IO is the first model capable of performing all 7 tasks on the GRIT benchmark and produces strong results across 16 diverse benchmarks like NYUv2-Depth, ImageNet, VQA2.0, OK-VQA, Swig, VizWizGround, BoolQ, and SciTail, with no task-specific fine-tuning. Code and demos for Unified-IO are available at: https://unified-io.allenai.org.